DOCUMENT RESUME

ED 077 876 SP 006 557

AUTHOR Gall, Meredith D.
TITLE The Problem of "Student Achievement" in Research on Teacher Effects.
INSTITUTION Far West Lab. for Educational Research and Development, San Francisco, Calif.
PUB DATE Feb 73
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana, February 1973

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Academic Achievement; *Achievement; *Achievement Gains; *Achievement Rating; *Curriculum Design; Motivation; *Teacher Behavior

ABSTRACT

This paper criticizes the work of Barak Rosenshine on the effects of teachers on student achievement. The author cautions against accepting Rosenshine's generalizations on teaching techniques. He makes the specific criticisms that Rosenshine a) did not operationally define student achievement, b) did not assess the validity of the student achievement measures used in his work, c) did not determine whether the achievement measures were appropriate to the students sampled, d) did not determine whether or not achievement measures were related to the curriculum objectives of the teachers, and e) combined the results of various studies without examining the relationships among them. Further, the author makes recommendations for future research on the topic. (JB)

ED 077876

THE PROBLEM OF "STUDENT ACHIEVEMENT" IN RESEARCH ON TEACHER EFFECTS

Meredith D. Gall

Far West Laboratory for Educational Research and Development

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SP 006 557

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973.

FILMED FROM BEST AVAILABLE COPY

THE PROBLEM OF "STUDENT ACHIEVEMENT" IN RESEARCH ON TEACHER EFFECTS

Meredith D. Gall

Far West Laboratory for Educational Research and Development

INTRODUCTION

Recently Barak Rosenshine has written several influential reviews covering a group of about fifty investigations into the relationship between teacher behavior and student achievement (Rosenshine, 1971a; Rosenshine and Furst, 1971). The purpose of my paper is to critically examine certain aspects of these investigations and of Rosenshine's reviews of them. The fact that my colleagues (Flanders, 1973; Heath and Nielsen, 1973) and I have organized an AERA symposium to critically evaluate these reviews expresses our recognition of the substantial impact that Rosenshine's work has had on research in classroom teaching.

Some of this impact has been positive. For example, Rosenshine has brought to our attention a large number of research investigations, most of them fairly recent and not easily accessible, which were designed to yield new knowledge about the effect of particular teaching practices on student achievement. He has also sensitized us to the need to evaluate whether we are devoting too much effort in training teachers in particular strategies and techniques; and not enough effort in testing, through carefully controlled research, whether these strategies and techniques will help children learn better.

Rosenshine has even attempted to advance the field of teaching research by telling us, based on his reviews of the literature, which teaching techniques are probably effective and deserving of further exploration; and which techniques are of lesser merit because they are not supported by existing research. For example, in one review (Rosenshine and Furst, 1971) he concludes:

> "Of all the variables which have been investigated in process-product studies to date, five variables have strong support from correlational studies and six variables have less support but appear to deserve future study... At first glance, the above list of the strongest findings may appear to represent mere educational platitudes. Their value can be appreciated, however, only when they are compared to the behavioral characteristics, equally virtuous and "obvious," which have not shown significant or consistent relationships with achievement to date." (pp. 54-55.)

In another review (1971a), he states:

> "One type of study which might evolve from this book is a correlational study in which the investigator attempts to replicate the importance of some of the variables which have been significant correlates of student achievement in previous studies." (p. 12.)

It is primarily toward this type of generalization by Rosenshine that my critical remarks will be aimed.

I will start by observing that the reviewer of the research literature on teacher effects faces a complex task. He must apply scholarly expertise to the evaluation of how successfully each investigation dealt with the following three aspects of research design: 1) measurement of the teacher's behavior; 2) measurement of gain in student achievement; 3) development of a theory or rationale to explain observed relationships between variables resulting from 1) and 2) above. If any one of these design considerations is not evaluated properly, there is a likelihood that the researcher will reach unjustified and misleading conclusions. My analysis of Rosenshine's reviews indicates that he has critically weighed at least some of the problems in measuring teacher behavior in research on teacher effects, but that he has neglected the other two design considerations. For example, in the 1971 monograph Rosenshine devotes six pages (pp. 18-23) to problems involved in developing and using a teacher observation instrument; another four pages (pp. 42-45) are spent discussing a particular instrument, Flanders' Interaction Analysis system. However, in the same monograph I could not find a single paragraph which confronts the problems involved in developing and using measures of student achievement. I shall attempt to show that this type of omission leads to an unbalanced review and to conclusions that may be unjustified (they do not legitimately follow from the research findings) and misleading (they are likely to be misinterpreted by a person unsophisticated in this field of research). My criticisms will refer primarily to Rosenshine's monograph (1971a) since it is the most comprehensive and most recent.

ROSENSHINE'S DEFINITION OF STUDENT ACHIEVEMENT

A good first step in conducting a research project, or reviewing a body of research, is to define one's terms. Obviously, generalizations of the type, "Judging by the available research, this variable has not been shown to be a significant predictor of student achievement" (Rosenshine, 1971a, p. 71), which are common in his reviews, are meaningless unless we know what is meant by the term, "student achievement." I was unable to find a formal definition of this critical term in Rosenshine's monograph. However, the following statement is suggestive:

> "This review is limited to studies of teacher behavior and student achievement; the relationships between teacher behavior and other important outcomes of schooling are not

> reviewed in this book. Such outcomes, which are also encompassed under the term teacher effectiveness, include: student attitudes toward self, school, and the subject area; creativity; disposition to use the subject area in the future; and personal development outcomes such as social sensitivity, self confidence, responsibility, social competence, and carefully thought out personal goals..." (Rosenshine, 1971a, p. 13.)

This statement tells us what student achievement is not; it fails to tell us what it is, though.

In another paper Rosenshine (1971b) provides a more direct definition of student achievement by drawing distinctions between the student outcomes of achievement, attitudes, and personal development. He defines achievement as follows: "Achievement refers to knowledge of facts, and also to skills of cognitive processing such as the ability to interpret, summarize, and compare information" (p. 77). Presumably the same definition would apply to the use of the term in the monograph. However, given the critical importance of this term in interpreting Rosenshine's generalizations, even this definition is insufficiently precise. A definition of "student achievement" that is scientifically adequate would include a description of the operations by which this concept was measured. This type of definition can be deduced from an examination of the instruments used to measure student achievement in the fifty investigations reviewed by Rosenshine. They can be described briefly as follows: they are paper-and-pencil tests; some are widely used standardized tests, others were specially developed for purposes of the investigation; some measure a limited range of curriculum objectives, others measure a wide range. Using this information, we can construct a definition of student achievement which corresponds to Rosenshine's use of the term in his reviews: student achievement refers to acquisition of facts or skills of cognitive processing, as measured by paper-and-pencil performance tests, standardized or locally developed.

Assuming the above definition is valid for Rosenshine's monograph, it is perhaps understandable that he would limit his review to studies that investigated the narrow range of student behaviors implied by the definition. What is not so understandable is why he neglected to define this critical term operationally and why he would use a broad term such as "student achievement" to denote a rather limited sample of the total range of behaviors that can be learned by students. Whatever the reasons, Rosenshine's use of the term in his reviews has two unfortunate consequences in my opinion.

The first unfortunate consequence is that the broad, undefined concept of student achievement obscures important value problems in the field of teaching. Let me explain. Teaching techniques such as using student ideas, providing praise, and asking higher cognitive questions are valued by educators for various reasons. Within the context of Rosenshine's reviews, though, these techniques are given value only when there is evidence that they are related to gains in student achievement. Thus, valuing particular teaching techniques is contingent upon valuing

student achievement, as defined by Rosenshine. The question can be posed: who values student achievement as defined by Rosenshine?

Rosenshine answers this question in another paper (1971b) by stating that, "Academic achievement is by far the outcome measure most acceptable to the majority of parents, students, teachers, and educators" (pp. 77-78). Perhaps this statement is true for student (academic) achievement as a broad label, but what about the operational referents that underlie the term? Do most educators value equally acquisition of facts and cognitive processing skills? I believe that most educators value the latter objective, but the former is definitely a controversial objective of current American education. For example, Ebel (1972) has taken a strong stand in favor of knowledge acquisition as a curriculum focus, but humanist educators such as Rogers (1971) and Holt (1967) have strongly criticized knowledge acquisition as outmoded in our technologically fast-changing society. Also, there are educators who believe that computer-assisted and programmed instruction will increasingly take over the role of instilling knowledge and developing simple cognitive skills, thus freeing the teacher to pursue other objectives. From this perspective, it makes little sense to evaluate teaching practices against the criterion of paper-and-pencil achievement tests proposed by Rosenshine. Finally, there are people who would value the general referents of Rosenshine's concept of student achievement, as I interpreted it, but who would object strenuously to some of the tests used to measure these referents.

The point is, to value Rosenshine's generalizations about particular teaching techniques, it is necessary to value his conception of student achievement. Since the term is not defined, the unwary reader may make a valuation that does not reflect his true feelings. To see why this is so, let us consider Rosenshine's generalization about research on teacher use of student ideas: "Judging by the available research, this variable has not been shown to be a significant predictor of student achievement" (1971a, p. 71). The "majority of parents, students, teachers, and educators" who value academic achievement might well value teacher use of student ideas less after reading such a statement. But if they knew the operational referents for "student achievement" in this context, would they form the same opinion? Suppose Rosenshine had stated, "Judging by the available research, this variable has not been shown to be a significant predictor of student gain on a limited range of paper-and-pencil tests, some standardized and some locally developed with unknown reliability,[1] which primarily measure fact recall and/or cognitive processing skills." I tend to think this statement would evoke a quite different valuative response from the reader than the one evoked by Rosenshine's generalization.

---

[1] See Heath and Nielson (1973) for a description of reliability of tests used in the studies reviewed by Rosenshine.

Another problem with using a broad term (student achievement) to denote a rather limited class of student behaviors is that it results in over-generalizations. Rosenshine admits that there is a great deal of research on teacher effects that he did not include in his review. However, statements of the type, "Judging by the available research, this variable has not been shown to be a significant predictor of student achievement," suggests that all the pertinent research has been reviewed, which is clearly not the case. Even more slippage occurs in the Rosenshine and Furst review (1971) of the same fifty studies. Their opening remarks contain the following statements:

> "This review is an admission that we know very little about the relationship between classroom behavior and student gains... In the first section of this paper we discuss the limitations of our knowledge about teaching, and acknowledge that sufficient information is not available on the relationship between a teacher's behavior and student learning in the classroom to design adequate programs in teacher education. In the second section we discuss the major results of one of the more promising areas of research on teaching -- those studies which attempted to relate observed teacher classroom behaviors to measures of student achievement." (p. 37, underlining mine.)

Considering the import of these sweeping generalizations, it is unsettling that the authors use terms such as "student gains," "student learning," and "student achievement" so loosely. Clearly, "student gains" and "student learning" refer to a broader class of behaviors than the term "student achievement." It may be true that we lack an adequate knowledge base for designing effective teacher education programs, but Rosenshine and Furst would have had to review a great deal more research on teacher effects than they did to justify reaching such a conclusion. The uncritical or unsophisticated reader could accept this conclusion as valid, though, if he did not perceive the discrepancy between the class of behaviors referred to in the conclusion ("student learning in the classroom"), and the class of behaviors covered in the review (a limited range of paper-and-pencil performance tests measuring primarily fact recall and cognitive processing skills.)

MEASUREMENT OF STUDENT ACHIEVEMENT

The use of standardized or locally developed tests to measure gains in student achievement is a hazardous procedure. However, Rosenshine did not critically evaluate the studies covered by his reviews to determine the extent to which certain hazards were avoided. I have already pointed out that in the 1971 monograph Rosenshine devoted six pages to problems involved in developing and using teacher observation instruments, but not a single paragraph to problems involved in developing and using student achievement tests. This omission raises serious doubts in my mind concerning the validity of the generalizations which Rosenshine reached on the basis of his review. My purpose here is not to critically review the

studies themselves to determine how effectively the researchers measured student achievement, but to point out methodological flaws which, if present, would threaten the meaningfulness of the results.

If an achievement test is not appropriate for the aptitude level of the research sample, the range of gain scores will be artificially restricted. Consider the case of a slow student who has benefited greatly from a teacher's instruction. His achievement will not be accurately reflected by the test if it contains too many of the items that he cannot answer, and too few of the items that he can answer. A different problem can confront the bright student. If the test items are generally at an easy level of difficulty, he will do very well on the pretest. Since the posttest is usually an alternate form of the pretest, he will do very well on it, too, but not much better because there are too few items on which he can demonstrate his superior competence. (There may also be a regression effect, which would also depress his post-achievement score). In short, if the sample observed by the researcher contains many of these students, either bright or slow, the range of gain will be artificially restricted. Restriction in range of scores is undesirable in teacher effects research because it lowers the value of correlational coefficients, thus underestimating the relationship between a particular teaching variable and a particular student achievement variable. Rosenshine did not report in his reviews whether he checked the investigations for presence of this statistical problem.

Perhaps the chief hazard to be avoided in using achievement tests in teacher effects research is lack of consistency between the curriculum content measured by the test and the curriculum content taught by the teachers in the researcher's sample. If the two are not consistent, the observed relationships between the teaching variables and student achievement variables will be impossible to interpret. To illustrate, suppose that teacher use of student ideas has the effect of improving students' ability to think constructively about the curriculum content they study in class. Therefore, if a group of students has a teacher who makes extensive use of their ideas while they are learning content X, they will achieve a higher level of performance than a group of students whose teacher makes little or no use of this technique. However, this difference will only appear if the researcher uses a test of achievement that samples randomly from content X. If the test samples primarily from content Y, the difference will be washed out. Thus, high consistency between the teacher's curriculum content and the test's content (i.e. high content validity) is absolutely critical in order for a research project on teacher effects to generate meaningful data. The exception is the case where the teacher's instruction can be assumed to produce transfer effects. If the instructor is teaching for transfer, then consistency between test content and curriculum content is not important; instead, consistency of skill objectives is important.

Unfortunately, Rosenshine did not make this critical check on the content validity of the tests used in the studies from which he derived his conclusions about the effectiveness of particular teaching practices. The

nature of the achievement tests used in the studies, though, raises doubts about their content validity. Many of the fifty studies involved the use of standardized achievement batteries such as the Stanford Achievement Tests, Sequential Tests of Educational Progress, California Achievement Tests, and Wide Range Achievement Tests.[1] Anastasi (1968) presents evidence that these tests are primarily measures of general scholastic aptitude or intelligence rather than of curriculum-specific content. For example, she makes the following points:

> "An examination of the content of several current instruments classified as intelligence and as achievement tests, respectively, reveals close similarity of content. It has long been known, moreover, that intelligence tests correlate about as highly with achievement tests as different intelligence tests correlate with each other (Coleman & Cureton, 1954; Kelley, 1927, pp. 193-209). In some instances, in fact, the correlation between achievement and intelligence tests is as high as the reliability coefficients of each test." (pp. 392-393.)

> "Few batteries today are directed primarily toward testing in content areas... Most batteries, at both elementary and high school levels, combine skill testing with some specialized content coverage." (p. 396.)

Concerning the Sequential Tests of Educational Progress (STEP) used in five of the studies reviewed by Rosenshine, Anastasi states, "...the heavy reliance of STEP on broadly oriented items brings this battery very close to scholastic aptitude or intelligence tests." (p. 400.)

If Anastasi's points have merit, it appears that many of the studies reviewed by Rosenshine are actually investigations of the relationship between teacher behaviors and student aptitude or intelligence.[2] In

---

1 Thirty-three of the seventy studies derived from the basic group of fifty investigations reviewed by Rosenshine in the 1971 monograph used these tests to measure student achievement (analysis based on data provided by Rosenshine, 1971a, pp. 45-51).

2 I am aware that many of the tests used in these studies were not standardized, but were developed especially for the particular investigation. One hypothesis which I have entertained is that these tests match the teachers' objectives better than standardized tests because they were developed by or under the guidance of the investigator. Assuming this is true, I would then hypothesize more powerful relationships between teacher and student behaviors when locally developed tests were used because the relationships would not be washed out by lack of consistency between test objectives and teacher objectives.

another publication concerned with the stability of teacher effects, Rosenshine (1970) himself expresses an awareness of this problem:

> "Such tests [standardized achievement tests] may be inappropriate measures of the influence of teacher behavior because the items on the tests may not be relevant to the materials or skills taught in the classroom. In many cases, these tests may be measuring the aptitude of the learner or the pressure for academic achievement in the home rather than the influence of the teacher." (p. 652.)

It is puzzling that Rosenshine did not mention this problem in his reviews and discuss how it would affect interpretation of the fifty studies.

The problem of matching test and teacher is further complicated by the fact that different teachers emphasize different educational objectives, even when they are teaching the same curriculum. This was the situation found in Bellack's classic investigation (1966) of teacher-student interaction patterns. He designed his project with the intent of investigating variations in teaching method, holding subject matter constant between teachers. However, Bellack found much greater variation in teachers' choice of subject matter within a given set of curriculum materials than in their choice of teaching methods. The implication of this finding is that it may be quite difficult to locate a standardized achievement test, or to develop one, which will be appropriate for all the teachers included in one's research sample. Still another problem occurs if the teachers rely heavily on individualized instruction in which each student pursues a different curriculum objective. Rosenshine did not report whether he checked the studies included in his reviews to determine the possible presence of these measurement problems.

There is one section in his 1971 monograph where Rosenshine displays awareness of the problem of matching test and teacher (pages 196-200). He cites several studies which demonstrate that opportunity to learn the material covered by an achievement test is significantly correlated with scores on that test. This is evidence that there is variation in teachers' curriculum objectives and that this variation affects their students' performance on tests. Unfortunately, Rosenshine did not take the next step by discussing how these variations complicate the interpretation of data yielded by teacher effects research. He might have pointed out that this problem and others which I discussed above would lead to conservative errors in revealing teacher effects. That is, these problems, if present, would create underestimates of relationships between teacher and student variables. Thus, the significant relationships identified by Rosenshine might be stronger than the data suggest. By the same token, the nonsignificant relationships should not be viewed as conclusive. They may reflect the presense of measurement errors rather than an actual lack of relationship between a particular teacher behavior and student achievement criterion.

THE RELATIONSHIP BETWEEN STUDENT ACHIEVEMENT AND TEACHER BEHAVIOR

The purpose of this section is to examine Rosenshine's conceptualization of the relationship between student achievement variables and teacher behavior variables. Also, alternative conceptualizations will be considered in order to reveal his particular biases in interpreting the relationships obtained in teacher effects research to date.

First, it is necessary to distinguish between two types of correlational research that can be done on teacher effects. I call the first type "empirical-exploratory" research; others have called it "dust-bowl empiricism." The distinguishing feature of this type of research is that it is not guided by theory or explicit rationale. The investigator simply observes various teacher behaviors and administers various student achievement measures; the two sets of variables are then correlated to determine whether there are underlying relationships. Sometimes, but not always, variables included in the correlation matrix appear to have been arbitrarily selected. I will label the second type of investigation "hypothesis-testing" research. In this type of research the investigator's choice of teacher behaviors to be observed and student achievement measures to be administered is guided by hypotheses derived from formal or informal theory. Compared to empirical-exploratory research, hypothesis-testing research is guided by an explicit rationale, and the results can be interpreted in terms of that rationale.

It appears that the majority of the research surveyed by Rosenshine followed the empirical-exploratory rather than the hypothesis-testing model.[1] The investigation by Wright and Nuthall (1970) is typical. In their report they describe the investigation as "exploratory" (p. 478). Their selection of teaching variables was guided by empirical findings of earlier studies, although not completely, since factor analysis was used to eliminate some variables and to combine others into composite variables. The student achievement measure was developed especially for this investigation, and a total of 28 teacher behavior variables was correlated with it. Relationships between variables were sorted out by means of statistical significance criteria, and the investigators provided some post-hoc interpretation of the data.

What knowledge about teacher effects should a reviewer try to derive from the type of research represented by Wright and Nuthall's study? Rosenshine's approach was to combine the results of several studies investigating similar teacher behavior variables and then do what I call a statistical scorecard tally, typified by summary statements such as, "The cognitive clarity of a teacher's presentation has been studied in seven investigations in which student or observer ratings were used. The investigators

---

[1] Personal communication from Mark Nielson, who reviewed the original reports of the fifty studies.

used different descriptions of clarity...Significant results on at least one criterion measure were obtained in all seven studies. In those studies for which simple correlations were available, the significant correlations ranged from .37 to .71." (Rosenshine and Furst, 1971, p. 44.) It appears that the main goal of Rosenshine's reviews was to yield a list of teacher behavior variables that have correlated with student achievement measures consistently at a level of statistical significance.

How useful is such knowledge? Such knowledge may or may not be useful, depending on two conditions. First, the statistical trends identified by Rosenshine need to be interpretable. Knowing that teacher clarity is consistently related to student achievement is of little value unless we can formulate a reasonable explanation of the observed relationship. For example, how can we design a meaningful follow-up experiment to determine whether teacher clarity causes improved student achievement unless we have some understanding of the nature of the relationship? Rosenshine himself did not attempt to interpret the results of his statistical summaries. I suspect that interpretation may be quite difficult given the diversity of teacher behaviors,[1] the diversity of student achievement measures, and the diversity of teacher populations subsumed under each of Rosenshine's statistical trend analyses. Incidentally, the same problem of interpretation arises for the nonsignificant results reported by Rosenshine. Consider the summary statement, "Of the 11 studies which employed linear correlations in the study of an i/d ratio, two yielded significant results..., seven yielded positive but nonsignificant results... and two yielded small negative results..." (1971a, p. 83.) How useful is such knowledge? I contend that it is of little use in strengthening or weakening our confidence in particular teacher behavior variables since Rosenshine did not interpret each nonsignificant result to determine whether it could be attributed to a methodological problem such as small sample size or inappropriate student achievement measure.

The second condition for determining the usefulness of statistical trend knowledge involves the student achievement measures used in the studies that were reveiwed. The usefulness of the knowledge depends upon the value that we attribute to these measures. If teacher clarity was consistently related to educationally significant measures of student achievement, this knowledge certainly would be more useful and important than if it were consistently related to trivial measures. Since Rosenshine did not analyze the student achievement measures from this perspective, the educational significance of his statistical trends is still in question.

To summarize, Rosenshine's approach of making statistical trend tallies from empirical-exploratory research is superficial. It needs to be accompanied by thoughtful analysis and interpretation of observed relationships

---

[1] Heath and Nielson (1973) have provided detailed criticism of Rosenshine's procedure of grouping studies that investigated similar teacher behavior variables.

between the particular teacher behavior variables and student achievement variables that were measured. Thoughtful interpretation is particularly necessary when so many of the studies that were reviewed appear to have been unguided by an explicit rationale or theory.

To illustrate the need for careful interpretation, and incidentally the wrong inferences that can be drawn when a statistical tally is made unaccompanied by analysis of particular results, I refer again to Wright and Nuthall's study. One of the teacher behaviors investigated by them, and reviewed by Rosenshine, was teacher use of closed questions (requiring fact recall) and of open questions (requiring judgment, interpretation, prediction, etc.). In Table 3.9 (1971a, p. 122) Rosenshine reports that teacher frequency of closed questions correlated +.31 with residual student achievement, and frequency of open questions correlated -.08 with the same criterion. Perhaps due to an oversight, he neglected to report in his section on ratios of closed and open-ended questions (1971a, pp. 126-130) that Wright and Nuthall also computed these percentages: percentage of closed questions, r = +.46, and percentage of open questions, r = -.21 with residual student achievement. In the text he dismisses the results on frequency because they are not statistically significant:

> "No significant results were obtained for the frequency of factual questions in five studies (Harris and Serwer, 1966; Harris et al, 1968; Spaulding, 1965; Wright and Nuthall, 1970)... The classification of all questions into only two forms has not yielded consistent significant results. The non-significant results are puzzling. One would expect that the frequency of questions that encourage students 'to seek explanations, to reason, to solve problems' (Perkins, 1965) or the frequency of questions related to interpretation (Harris and Serwer, 1966; Harris et al, 1968) would be consistently related to achievement."
> (1971a, pp. 123-124.)

According to the last sentence, Rosenshine had a certain expectation concerning frequency of open questions and student achievement, which was not fulfilled by the research results. However, Rosenshine failed to ask himself a prior question: was there any reason to expect a relationship between these two constructs (frequency of open questions and student achievement), given the particular measures that were used in the studies? Let us examine the particular student achievement measure used by Wright and Nuthall. Their post-training achievement test consisted of "29 multiple-choice items" (p. 480), "limited to a particular set of educational objectives (knowledge of an elementary science topic)" (p. 489). Given this information, the observed relationships are easily interpreted: if a teacher wants pupils to acquire facts, it may be helpful to ask a high percentage of closed questions (r = +.46 with the criterion); open questions should be inconsequential, or perhaps detrimental, if they limit opportunity to ask closed questions or if they sidetrack students from drilling on facts in preparation for the criterion test. Even if we share Rosenshine's expectation that "frequency of (open) questions... would be consistently related to

achievement," I doubt that anyone would expect this behavior to be related to the particular achievement measure used by Wright and Nuthall.

The above discussion is not meant to call into question Rosenshine's contribution to the field of research on teaching. Undeniably he has moved the field perceptibly forward through his reviews. The point of my criticisms is to indicate the need to view the results of his statistical trend tallies with caution until they can be substantiated by careful analysis and interpretation of the relationships summarized by them. As the tallies stand now, they represent too shallow an approach. What is needed are additional reviews which take a more in-depth look at the findings.

A BROADER VIEW OF STUDENT ACHIEVEMENT

The goal of research on teacher effects, and of reviews of this research, should be to increase our understanding of how teachers make a difference under prevailing school conditions and how they can make a difference under new school arrangements not yet widely used (e.g. individualized instruction; the open classroom). To achieve this goal, we need to develop a better understanding of our value commitments concerning student learning, how these commitments influence what we have chosen to study or review, and what we can expect to find as a result of what we have chosen to study or review. The first step is to map the broad range of student behaviors that conceivably could be influenced by the teacher.

Perhaps the most comprehensive list of student behaviors is provided by the three taxonomies of Benjamin Bloom and his associates. A much simpler, but nevertheless useful classification is given by Rosenshine (1971b). He uses the broad term "student growth" to denote three types of student behavior that can be learned:

> 'Achievement' refers to knowledge of facts, and also to skills of cognitive processing such as the ability to interpret, summarize, and compare information.
>
> 'Attitudes' refers to a variety of measures which may or may not be interrelated: attitudes toward self, school, or subject areas; out of school activities such as browsing in a library or going on nature walks; and dispositions to use cognitive skills in future activities.
>
> 'Personal development' refers to a variety of outcomes such as self-confidence, ability to persist in difficult tasks, disposition to inquire into new problems, assumption of personal responsibility, ability to make reasoned choices, curiosity, and development of independence. (p. 77.)

It is also necessary to distinguish two ways in which these student behaviors and attitudes can be influenced by a teacher. The type of teacher influence which comes to mind most readily is facilitation of learning of these behaviors and attitudes. Less attention has been given to a type of influence which I will call "elicitation of student performance." For example, many students are capable of responding actively (e.g. answering questions) or passively (e.g. listening) in the classroom; they have already learned these sets of behaviors. The teacher can exert influence by eliciting the performance of one, the other, or both of these sets of behaviors. To be more specific, suppose a teacher asks a student for his opinion; the student already knows his opinion and simply articulates it. No learning may have occurred, but the teacher has exerted influence by eliciting a certain type of performance. The concept of eliciting performance of a behavior may be particularly helpful in understanding the relationship between teacher behavior and students' mood states. Children have undoubtedly learned to be happy, sad, whimsical, or angry before they enter school. However, the teacher can still exert influence by using techniques which elicit various of these mood states.

Teacher elicitation of student performance may be valued for two reasons. First, we may value the elicitation in its own right. For example, we can value teacher techniques which elicit a high percentage of student talk in class discussions, because we value student talk. Another reason why we can value these techniques is that they elicit a high percentage of student talk, which in turn facilitates learning or elicitation of other behaviors which we value. In other words, we can value a particular student performance as an end in itself or as a means to an end.

How can these distinctions be used to clarify our thinking about research on teacher effects? When a researcher asks himself the question, "How might this particular teaching behavior affect student behavior?", he can use these distinctions to consider a broad range of student behaviors and types of teacher influence. If he chooses only to investigate the teacher's influence as a facilitator of student learning on particular achievement tests, this may reflect a value commitment which, if present, should be clearly noted in his report of findings. Also, the researcher needs to ask himself whether he has selected particular student measures only because he values them or also because he has an explicit rationale to explain why they measure an appropriate set of student behaviors, given the teacher behaviors that are being observed.

My emphasis on value commitments in research on teacher effects has implications beyond the selection of particular teacher and student behaviors to be observed. The teachers who are subjects in this research have their own value commitments which can affect the researcher's data. I expect to see many experimental studies in the future of the type recommended by Rosenshine (1971a, p. 12): train teachers to use a particular technique and then test for effects on their students. The researcher needs to determine whether the teachers value the technique, the changes in their behavior, and the presumed effect(s) of the technique on students. If teachers do not value these things, they may use the technique on demand

(i.e. when the researcher is observing), but not otherwise and not in such a way as to produce the desired effect on students.

SUMMARY

The points that I have made in this paper can be summarized under the general criticism that Rosenshine did not consider a number of problems involved in measuring student achievement in research on teacher effects. Thus, although Rosenshine's reviews are a landmark in our field, his generalizations concerning the demonstrated effectiveness of particular teaching techniques should be viewed with caution until further analysis and interpretation are made of the studies which he reviewed. My specific criticisms are that Rosenshine:

1. did not provide an operational definition of the term "student achievement,"

2. did not evaluate the educational worth of student achievement measures used in the research,

3. did not analyze the achievement tests to determine whether they were appropriate measures for the students used in the research,

4. did not analyze the achievement tests to determine whether they sampled adequately the curriculum objectives taught by the teachers,

5. combined the results of several studies into a scorecard tally of statistical significance without also analyzing and interpreting the meaning of the observed relationships.

It is easy to take these criticisms of Rosenshine's reviews and turn them into recommendations for future reviews of teacher effects research and for the design of such studies. I have two additional recommendations to make concerning the latter task:

1. If a researcher intends to evaluate the effectiveness of a particular teaching technique, he would do well to consider the total range of student behaviors that might be affected by the technique. In addition, he should consider the possibility that teacher use of the tecnnique might elicit student performance rather than affect student learning of certain behaviors.

2. Although it is easier said then done, I strongly encourage the use of psychological theory, or at least of an explicit rationale, to guide future investigations in this area.

REFERENCES

Anastasi, Anne. Psychological testing. (Third edition). New York; Macmillan & Company, Inc., 1968.

Bellack, A.A. and others. The language of the classroom. New York; Teachers College Press, 1966.

Ebel, Robert L. What are schools for? Phi Delta Kappan, 1972, 54, 3-7.

Flanders, Ned A. Knowledge about teacher effectiveness. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973.

Heath, Robert W. and Nielson, Mark. The myth of performance-based teacher education. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973.

Holt, John. How children learn. New York: Pitman, 1967.

Rogers, Carl. Forget you are a teacher. Instructor, 1971, 81, 65-66.

Rosenshine, Barak. Teaching behaviours and student achievement. London: National Foundation for Educational Research in England and Wales, 1971a.

Rosenshine, Barak. New directions for research on teaching. In How teachers make a difference. Washington: U.S. Government Printing Office, 1971b. (Catalog No. HE 5.258:58044.)

Rosenshine, Barak. The stability of teacher effects upon student achievement. Review of Educational Research, 1970, 40, 647-662.

Rosenshine, Barak and Furst, Norma. Research in teacher performance criteria. In B.O. Smith (ed.) Research in teacher education: A symposium. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Wright, C.J. and Nuthall, G. Relationships between teacher behaviors and pupil achievement in three elementary science lessons. American Educational Research Journal, 1970, 7, 477-493.